

SEGMENTATION AND ESTIMATION OF CHANGE-POINT MODELS

XIAO FANG, JIAN LI AND DAVID SIEGMUND

*Chinese University of Hong Kong, Adobe Systems and Stanford
University*

Abstract

To segment a sequence of independent random variables at an unknown number of change-points, we introduce new procedures that are based on thresholding the likelihood ratio statistic. We also study confidence regions based on the likelihood ratio statistic for the change-points and joint confidence regions for the change-points and the parameter values. Applications to segment an array CGH analysis of the BT474 cell line are discussed.

1 Introduction

Diverse applications of change-point problems have led to recent interest in segmentation of models involving multiple change-points. A model having some direct applicability and additional theoretical interest for the insights it provides is as follows. Let X_1, X_2, \dots, X_m be independent and normally distributed with variances equal to 1. Assume that there exist $M \geq 0$ and integers $0 = \tau_0 < \tau_1 < \dots < \tau_M < \tau_{M+1} = m$ such that the mean μ_i of X_i is a step function with constant values on each of the intervals $(\tau_{i-1}, \tau_i]$, but different values on adjacent intervals. Segmentation amounts to determining the value of M , the τ_i and perhaps also the μ_i . Because of the computational difficulty of sorting through all possible partitions of $[1, m]$ to find the change-points when m is large, there have often been different algorithms for suggesting a set of candidate change-points τ_i and for determining which of those possible sets is “correct.” For example, one might use a dynamic programming algorithm to propose a relatively small set of possible M and $\tau_i, 1 \leq i \leq M$, then use a statistical procedure to determine a final choice from those suggested in the first stage of analysis. For recent reviews imbedded in otherwise original research articles see Frick, Munk and Sieling (2014)

and Fryzlewicz (2014). Another recent paper that provides motivation for our research is Niu and Zhang (2012), who also emphasize tentative selection of several sets of candidate change-points followed by model selection to make the final choice. The procedures suggested below could also form part of a two-stage procedure, but here we consider in detail only a single stage.

Substantial motivation for recent research has been copy number variation in genetics, and reference to this problem can help us determine interesting sample sizes and values for parameters in our numerical examples. In particular, the sample size m can be quite large, and M can be large in an absolute sense, while still small compared to m . An early genomic application involves Bernoulli variables, which at a genomic location are labeled 0 or 1 according as the DNA letter at that location is A or T, or is C or G. Since a CG “rich” region is an indication of the presence of a gene or genes, it may be useful to segment a genome or part of a genome into regions of relatively low or high CG content. See, for example, Churchill (1989) who used a Hidden Markov Model, or Elhaik, Graur, and Josić (2010).

A variety of other examples motivated by particular scientific experiments is given by Du, Kao and Kou (2016). In particular they describe examples where several consecutive changes are expected to have the same sign and where the pattern of change-points arises from a hidden Markov model.

Our goal is to consider two new methods, which in principle combine the two steps described above. We also suggest a method that, for a given value of M , provides an approximate likelihood ratio confidence region for the sets of mean values μ_i and the change-points τ_j , or for just the change-points with the mean values as nuisance parameters. We also compare our approximate joint confidence region with the very interesting suggestion of Frick, Munk and Sieling (2014).

Let $S_j = \sum_1^j X_i$. Consider the generalized likelihood ratio statistic for testing the hypothesis $M = 0$ against $M = 1$: $\max_j |S_j - jS_m/m|/[j(1 - j/m)]^{1/2}$. This statistic is the basis of the binary segmentation suggestion of Vostrikova (1981), which is a “top down” procedure, in the sense that one tests all the data to determine if there is *at least* one change-point and iterates the procedure in the subset immediately to the “left” and to the “right” of the most recently detected change-points. How one should choose a threshold at each stage is unclear. For a given threshold, the false positive probability for this statistic is (numerically) subadditive in m , so use of the same threshold at each iteration can lead to an uncontrolled false positive rate when the number of changes is large. For applications of this general approach to copy number variation based on a different class of statistics that

seem to reflect better the science motivating the problem, see Olshen *et al.* (2004) and Zhang *et al.* (2010).

In this paper we study "bottom up" procedures motivated by the observation that in the presence of multiple change-points, it seems appropriate to compare a candidate change-point at j to an appropriate local background (i, k) , where $i < j < k$. To that end, consider $\max_{i,j,k} |Z_{i,j,k}|$, where for $i < j < k$:

$$Z_{i,j,k} = [S_j - S_i - (j-i)(S_k - S_i)/(k-i)] / [(j-i)(1 - (j-i)/(k-i))]^{1/2}. \quad (1)$$

Given an appropriate threshold, because of local correlations, this statistic will typically exceed the threshold at a number of neighboring values of i, j, k . Two methods for sorting these values are discussed below.

Motivation for our study are provided by related statistics found in Fryzlewicz (2014), whose "wild binary segmentation" uses a random subset of integers i, k , then compares $\max_j |Z_{i,j,k}|$ to a threshold, and in Niu and Zhang (2012), whose statistic is similar to (1), although the background is required to be symmetric about j .

Our first theoretical result is an approximation for the tail of the distribution of (1) (and related statistics introduced below) when there are no change-points. Thresholding (1) produces a set of candidate change-points j , with multiple backgrounds (i, k) . We require that the background for one candidate change-point j not overlap another candidate change-point j' in the sense that if $j < j'$, the corresponding backgrounds should satisfy $k \leq j'$ and $i' \geq j$. This can be accomplished by sequentially re-evaluating candidate change-points until they satisfy the constraint. This re-evaluation may require that we alter a putative change-point or even remove it from the list. An approach that requires very little and usually no re-evaluation of candidate change-points is to select the shortest of the possible backgrounds (i, k) from among those for which $Z_{i,j,k}$ exceeds the required threshold. If there is a tie for the shortest value of $k - i$, we choose the one with the largest value of $|Z_{i,j,k}|$. Some experimentation indicates that selection based on the largest value of the statistic rarely leads to significant differences from selection based on the shortest background. Still another possibility is a pseudo-sequential procedure where we initially set $i = 0$, find the smallest $k > i + m_0$, $m_0 \geq 1$, such that $\max_{i < j < k} |Z_{i,j,k}|$ is above an appropriate threshold, set j_1 equal to the largest such j or the maximizing value of j , then set $i = j_1$ and iterate. See the unpublished Stanford Ph. D. thesis of E.S. Venkatraman for an early discussion of a similar idea.

As mentioned above, a statistic suggested by Niu and Zhang (2012) uses

the local maxima with respect to j of

$$Z_{j,h} = |[(S_{j+h} - S_j) - (S_j - S_{j-h})]|/(2h)^{1/2}, \quad (2)$$

where h is a parameter to be chosen. This statistic also compares the data at a putative change-point to the local background, which is chosen to be symmetric about the putative change-point. There is no obvious choice for h without prior knowledge of the data, so Niu and Zhang suggest maximizing (2) over a finite number of values of h . For their applications to copy number variation, they suggest 3 values, 10, 20, and 30.

When there is no change-point, well known results allow us to compute a good approximation to the maximum of (2) for fixed h , but corresponding results for maxima over both j and h do not seem to be available, since the increments obtained by perturbing j and h by small amounts are not independent. However, an approximation computed as if they were independent appears to be reasonably accurate, as we indicate below. An alternative that gives similar results is to use a suitably smoothed statistic, which for simplicity we describe in continuous time.

Suppose W_t is Brownian motion with unit variance per unit time and drift parameters forming a step function as indicated above. Let $X_t = \int_{-\infty}^{\infty} f(t-s)dW(s)$, for a suitable smooth function f . For illustrative purposes below we use a Gaussian kernel, usually with unit variance. Then let

$$Z_{t,h} = |[\int_t^{t+h} X_s ds - \int_{t-h}^t X_s ds]|/\sigma(h), \quad (3)$$

where $\sigma(h)$ is the standard deviation of the numerator. In practice both the definition of X_t and the statistic (3) must be defined by summation, but for a reasonably smooth kernel f , the difference between the continuous and the discrete model seems unimportant.

Remarks. (i) There are advantages to the formulation in (3) beyond the fact that we can rigorously study its maximum over suitable ranges of both t and h (cf. (2.4) below). Regarding the problem discussed in this paper, calculations not shown here suggest that using (3) sacrifices very little power compared to (2). Perhaps more importantly it may prove useful in other problems, e.g., problems in several dimensional time, where it is possible to analyze smooth random fields designed to detect signals having a variety of shapes, or even in one dimension if one wants to incorporate dependence into a model for the residuals.

(ii) For some genomic applications, e.g., for inherited copy number changes, the signal to be detected extends over a relatively short range in the form

of a “peak,” where one change is followed by a second, nearby change in the opposite direction. For problems of this form it may be sensible to use statistics adapted to the expected shape of the signals. We briefly return to this point in Section 8.

(iii) We have assumed the variance is known. Since the sequences are usually long, the variance can in most cases be estimated sufficiently accurately that it can effectively be assumed known. In simulations to investigate this issue, we have used one half the average of the squared differences of consecutive observation as an estimator, so the heterogeneity arising from the change-points does not substantially inflate our variance estimate.

2 Approximate p-values

Assume that the μ_i are identically constant. Consider display (1) maximized over $0 \leq i < j < k \leq m$ and $m_0 \leq j - i, k - j \leq m_1$. (An alternative constraint on the size of the background, which yields a slightly different approximation, is $k - i \leq m_1$.) Then for large b and m

$$\mathbb{P}\{\max_{i < j < k} |Z_{i,j,k}| \geq b\} \sim .25b^5 \varphi(b) \\ \times \sum_{m_0}^{m_1} \sum_{m_0}^{m_1} \frac{(m - u - v)}{uv(u + v)} \nu[b(\frac{u}{v(u + v)})^{1/2}] \nu[b(\frac{v}{u(u + v)})^{1/2}] \nu[b(\frac{u + v}{uv})^{1/2}]. \quad (4)$$

Here φ is the standard normal probability density function. The function ν is defined, e.g., in Siegmund and Yakir (2007) p. 107 and given to a simple approximation by the equation

$$\nu(x) = (\Phi(y) - 1/2)/[y(\Phi(y) + \varphi(y))],$$

where $y = x/2$, and Φ is the standard normal distribution function.

For applications where very short intervals between change-points can occur, we often want to take $m_0 = 1$. There usually is considerable flexibility in choosing m_1 . Relatively small values of m_1 serve two purposes. They can be used to minimize detection of many small jumps, which may themselves reflect experimental artifacts leading to drift in the underlying distributions (cf. Olshen *et al.* (2004), Zhang *et al.* (2010)); and they speed up what may otherwise be time consuming computations for large values of m .

For the pseudo-online procedure, we have a similar approximation:

$$\begin{aligned} & \mathbb{P}\{\max_{0 < j < k \leq m} |Z_{0,j,k}| \geq b\} \sim .5b^3\varphi(b) \\ & \times \sum_{1 < k \leq m} \sum_{0 < j < k} j^{-2} \nu[b(((k-j)/(jk))^{1/2})] \nu[b(k/(j(k-j)))^{1/2}]. \end{aligned} \quad (5)$$

This procedure has the advantages that it is much easier to implement than (1). Moreover, for any nominal significance level, the corresponding threshold is lower, since it involves maximization over two indices rather than three. It has the disadvantages that we do not have as strong a theoretical guarantee that the false positive error probabilities are controlled as effectively as when we use (1), and estimation of the change-points may not be as accurate.

Similarly, an approximation to the tail probability of the maximum over j and h of (2), calculated on the assumption that the local increments obtained from perturbations of j and h are independent is given by

$$\begin{aligned} & \mathbb{P}\{\max_{0 < t < m, 0 < h < \min(t, m-t)} |Z_{t,h}| \geq b\} \\ & \sim 1.5mb^3\varphi(b) \sum_h \nu[b(3/h)^{1/2}] \nu[b(1/h)^{1/2}]/h^2. \end{aligned} \quad (6)$$

Now consider (3). Let $f_2 = f * f$, $F(x) = \int_{-\infty}^x f(y)dy$, and $F_2 = F * F$. Let $R(u) = [2F_2(u) - 1 - 2(F_2(2u) - F_2(u))]/\sigma^2(u)$, $d_{11}(u) = 2(3f_2(0) - 4f_2(u) + f_2(2u))/\sigma^2(u)$ and $d_{22}(u) = 2(f_2(0) - f_2(2u))/\sigma^2(u) - R^2(u)$. Finally, let $D(u) = d_{11}(u)d_{22}(u)$.

Then if μ_i is identically constant,

$$\begin{aligned} & \mathbb{P}\{\max_{m_0 \leq t \leq m_2, h_0 \leq h \leq h_1} |Z_{t,h}| \geq b\} \\ & \sim 2m(2\pi)^{-1}b\varphi(b) \int_{h_0}^{h_1} D^{1/2}(u)du. \end{aligned} \quad (7)$$

Proofs of (4) - (7) are deferred to an appendix.

Remarks. The approximation (4) exerts some control over the existence of false positive errors, even if there are change-points. Suppose, for example, that there are change-points at $j_1 < j_2 < \dots < j_M$ and put $j_0 = 0$, $j_{M+1} = m$. Then the maximum in (4) is stochastically larger than $\max_{0 \leq r < M} \max_{j_r \leq i < j < k \leq j_{r+1}} |Z_{i,j,k}|$, so the probability of a false positive satisfying these constraints would be (approximately) dominated by the right hand side of (4).

3 Simple Numerical Examples

In this section we begin with evidence that the false positive approximation given by (4) in the preceding section is reasonably accurate. Table 1 compares the approximation (4) with simulated values obtained from 2000 repetitions.

Table 1: Approximation (4). Simulated values based on 2000 repetitions.

b	m	m_0	m_1	p_{Approx}	Monte Carlo
4.71	500	1	50	0.05	0.053
4.60	500	1	100	0.10	0.103
4.77	500	1	100	0.05	0.054
4.71	500	3	100	0.05	0.043
4.45	500	3	50	0.10	0.108
4.68	300	1	299	0.050	0.049
4.76	400	1	399	0.051	0.049
4.83	500	1	499	0.050	0.047

Some numerical experimentation, not reported here in detail, suggests that the approximations (5) and (6) are also adequate. For example, for $m = 500$, and $b = 4.47$, the approximation (6) gives 0.0508, while a 25000 repetition simulation experiment gives the 0.05 threshold as 4.44. For (5) the threshold $b = 4.34$ yields the probability 0.051; for that threshold simulations (25000 repetitions) give the probability 0.045.

We now turn to a consideration of the power to detect change-points. In all cases we use a threshold that controls the global false detection probability to 0.05. For the statistics (1) and (2), we use the approximations (4) and (6), respectively. For the Niu-Zhang (NZ) statistic, we use three values of h , 10, 25, and 50; and we use a threshold so that the probability that the sum over these three values of h of the probability that the maximum over j of (2) equals 0.05. For the distribution of the maximum over j of (2), we use the asymptotic approximation discussed in Siegmund and Yakir (2007), p. 112, which was originally designed for a different problem, but is asymptotically equivalent to the problem at hand. (The correlation between two values of (2) separated by a j units is $(1 - 3|j|/2h)^+$. The local rate of decay, $3/2h$, should be equated to $\beta\Delta$, which is the local rate of decay of the Ornstein-Uhlenbeck correlation function, $\exp(-\beta\Delta|j|)$, for the process discussed by Siegmund and Yakir (2007)).

For the first calculation, we assume that there is a change-point of size δ at j^* , which is bracketed by $i^* < j^* < k^*$, where there are no other change-points; and we compare the marginal power of the three procedures. By marginal power we mean the probability to find the statistic above the appropriate threshold under the assumption that the statistic is focused on the true change-point and an appropriate background. For the statistic (1) it means that we use the true values of i, j, k . For (2) it means that $j = j^*$ and $h = \min(k^* - j^*, j^* - i^*)$. Hence if $k^* - j^*$ is about equal to $j^* - i^*$, which is effectively an assumption motivating the use of (2), then we expect to find that (2) has greater power than (1) by virtue of its lower threshold. For (2) for values of $\min(j^* - i^*, k^* - j^*)$ equal or slightly greater than the one of the three designated values of h , we compute the marginal probability of exceeding the threshold using that value of h . It is unclear how we should evaluate the marginal power for values of $\min(j - i, k - j)$ slightly less than one of these values of h , since the probability of being above the threshold for that value of h will depend on the value of μ outside the interval (j, k) .

In Table 2 $m_0 = 1, m_1 = 200, m = 10000$. The threshold values are $b = 5.49$ for the statistic (1), $b = 5.12$ for (2), and $b = 4.67$ for the NZ statistic with the three values $h = 10, 25, 50$. (Since the tail probabilities of (2) decrease for increasing h , one could add a fourth value, $h = 100$, and the threshold would increase only to 4.70.) Without loss of generality we set $i^* = 0$. We see from the table that when $k^* - j^*$ is about equal to $j^* - i^*$ and about equal to one of the three values of h , the marginal power of the NZ statistic is larger. In cases when the differences $k^* - j^*$ and $j^* - i^*$ are about equal to each other, but differ from one of the nominal values of h , the adaptive version of (2) can have more power; and when these differences are unequal, (1) can have more power. The qualitative effects correspond to our intuition.

Table 3 makes a similar comparison of (1) and (2) for an example that seems typical in studies of copy number variation. The example was originally suggested by Olshen *et al.* (2004) and used by Zhang and Siegmund (2007) as well as Frick, Munk and Sieling (2014). The number of observations is 497. Six changepoints occur at $j = 138, 225, 242, 299, 308, 332$. The sizes of the changes are 0.26, 0.99, -1.60, 0.69, -0.85, 0.53, respectively. We take the standard deviation of the observations to be 0.4, which is larger than in previous studies. For our statistics we take $m_0 = 1, m_1 = m = 497$. The threshold for (1) is 4.81; that for (2) is 4.47. The value for δ is given in standardized units.

Because of the unequal spacing of the change-points, the marginal power of (1) is at least as large as that of (2) for each change-point; it can be

Table 2: Marginal power I

δ	$j^* - i^*$	$k^* - i^*$	Power(1)	Power(2)	Power(NZ)
2.5	10	20 (35)	0.54 (0.88)	0.68	0.82
2.5	12	24 (30)	0.74 (0.89)	0.84	0.82
2.0	15	30 (40)	0.49 (0.74)	0.64	0.42
1.5	25	50 (65)	0.43 (0.70)	0.57	0.74
1.5	33	66 (90)	0.73 (0.91)	0.83	0.74
1.0	50	100 (150)	0.31 (0.61)	0.44	0.63
1.0	60	120 (175)	0.49 (0.79)	0.64	0.63
0.9	75	150 (200)	0.51 (0.75)	0.65	0.43

Table 3: Marginal power II. Six change-points.

δ	i^*	j^*	k^*	Power(1)	Power(2)	Local Power
0.65	0	138	225	0.47	0.43	0.69
2.5	138	225	242	1.00	1.00	1.00
4.0	225	242	299	1.00	1.00	1.00
1.73	242	299	308	0.50	0.21	0.65
2.13	299	308	332	0.73	0.52	0.82
1.33	308	332	497	0.90	0.55	0.95

substantially larger when one of the spacings is small and the other larger. Other examples with 10 change-points chosen at random from the interval $[1, 500]$ produce roughly similar results, although the marginal power of (2) can be larger in some instances.

The final column of Table 3 contains for the statistic (1) what we have called the local power, which gives the probability of detection for *some* (i, j, k) in a neighborhood of (i^*, j^*, k^*) . See Table 13 for simulations of these probabilities and the Appendix for a discussion of the analytic approximation used.

4 Algorithms and Simulations

Although the marginal power given in the preceding section provides some insight into these procedures, we want to consider the effect of using “local”

statistics in segmenting multiple change-points. In this section we report the result of simulations to compare a number of different segmentation procedures, with emphasis on their efficiency to detect change-points without an excessive number of false positive errors.

Table 4 shows the outcomes of some individual simulations, for different locations of change-points and sizes of the jumps, which are chosen to illustrate some of the observations we have already made. In particular we would like to consider more carefully the apparent weakness of (2) when the distance between consecutive change-points is small and consider the value of the pseudo-sequential procedure (called Seq in the Table). The sample size is $m = 300$, and there are four change-points. The initial mean value is 0; the locations of the change-points and mean values after each change are given in the Table. A cell entry of 0 indicates failure to detect the relevant change-point. An asterisk calls attention to a detection that misses the true change-point by more than 10 observations.

In Table 4 our implementation of (1) was to choose the values j by minimizing the associated length of the background $k-i$ from among those values of $|Z_{i,j,k}|$ exceeding the threshold. If there was a tie in the minimum values of $k-i$, we chose the values i, j, k having the largest value of $|Z_{i,j,k}|$. If necessary, we enforced the condition mentioned above that the backgrounds not overlap. The other possibility mentioned above, to choose the largest value of the statistic, but then enforce the no overlap condition for the background values, rarely leads to a substantial difference. For example, for the first three groups in Table 4, there would be no changes in the the number of change-points detected and negligible changes in the estimated locations of the change-points. Additional experimentation not reported here produces similar results. Although the algorithm based on the maximum Z -value appears to make more complete use of the data and may have advantages, for example, if one wants to estimate the magnitudes of the changes, we have used the algorithm based on the minimum background length, which is slightly simpler since it often satisfies the no overlap condition without any iterations.

It seems clear from the results in Tables 3 and 4 that the statistics requiring a symmetric background do not do well in detecting very short intervals of change. A more subtle observation from Table 4 is that their estimate of the location of the change can be biased by their use of a longer than necessary background in order to maintain the required symmetry. They also have trouble when several consecutive changes are all positive or all negative. That said, their thresholds are substantially lower than the corresponding threshold of (1), which allows them some successful detections

Table 4: Examples of segmentations: $m = 300$, $b_{\text{Seq}} = 4.21$, $b_{(1)} = 4.68$, $b_{(2)} = 4.34$. The initial mean value is 0. Locations of change-points and mean values after the change are as indicated.

Procedure/Parameters	138, 0.75	199,2.5	208, 0.25	232, 1.5
Seq	141	198	206	232
(1)	141	198	207	232
(2)	140	0	207	240
Seq	142	198	206	230
(1)	141	198	207	231
(2)	138	0	207	231
Seq	133	0	206	230
(1)	134	0	207	228
(2)	113*	0	207	226
	25, 1.5	40, 0.1	175, -0.9	225,-0.1
Seq	26	37	173	227
(1)	26	41	173	226
(2)	24	49	173	227
Seq	23	38	170	0
(1)	23	39	170	0
(2)	24	0	170	0
Seq	23	41	188*	224
(1)	0	41	188*	0
(2)	0	41	0	224
	50,2.5	55, 0.2	200, 1.0	250, -0.1
Seq	48	54	210	248
(1)	49	54	207	247
(2)	49	0	203	247
Seq	48	53	193	247
(1)	48	54	193	247
(2)	44	0	196	250

that are missed by (1). Overall, the pseudo-sequential procedure seems best for these few examples. Although the problem does not appear in Table 4, the pseudo-sequential procedure tends to have slightly more false positive errors, because of the way the same threshold is reused with each detection.

Table 5 provides the outcomes of 500 simulations for detecting M change-points randomly located from 0 to 500 with the sizes of the changes taken to be normally distributed with mean value 2.5ξ , where the values ξ are independently ± 1 with probability $1/2$ and variance 0.5 . The first method uses the maximum $\max |Z_{i,j,k}|$, with segmentation based on the smallest value of $k - i$ of those statistics that exceed the 0.05 level threshold, $b_1 = 4.83$; the second is the sequential version described above, with the threshold 4.33 ; the third is the Wild Binary Segmentation procedure of Fryzlewicz (2014) with threshold $b_{\text{WBS}} = 4.565$. (This threshold is close to, but slightly different from the value $1.3[2\log(m)]^{1/2} = 4.58$ recommended by Fryzlewicz, which was presumably determined by numerical experimentation. Our threshold is the 0.15 significance threshold according to the approximation (4).)

In an early stage of this research we also considered a number of other algorithms that we found not to be competitive with those considered in Table 5.

For the reasons given above we eliminated from consideration the Niu-Zhang algorithm based on (2).

We tried the classical binary segmentation procedure, which is a top down iterative procedure that after M detections, tests the $M + 1$ intervals between putative change-points searching for an additional change-point. Since the significance level for a fixed threshold is sublinear in the length of the interval of observation, a fixed threshold may lead to an excess of false positive errors if the number of change-points is large. In our simulations we were unable to find a threshold that balanced the over/under detections to our satisfaction, and hence we did not include this procedure in Table 5.

We also tried, then eliminated the algorithm of Frick, Munk and Sieling (2014), which our early simulations using their default parameters found not to be competitive. (See, however, the Discussion in Section 8.)

Still another algorithm we tried was that of Du, Kao and Kou (2016). Their Bayesian algorithm is designed for problems where a normal mean and normal variance can both change. For our formulation, where the variance is assumed not to change, some limited simulations suggest that their procedure makes too many false positive and false negative errors. Our likelihood ratio procedure can in principle be adapted for their problem when both mean and variance may change, but to avoid overburdening the current paper we are leaving that problem for future research.

Remark. Although the pseudo-sequential procedure also repeatedly uses the same threshold, it does so with smaller and smaller remaining numbers of observations, so its false positive error rate does not increase substantially, even when the number of change-points is large. On occasion the procedure will provide detection at values of j less than and k greater than a true change-point, with the result that a second change may be detected immediately. This problem can usually be eliminated by enforcing a no overlap condition.

The algorithm based on (1) requires of order m^3 computations, hence can be slow for large values of m . A version with a relatively small value of m_1 , which seemed appealing because it would be faster to evaluate and would exert some control over the detection of small changes, which often represent artifacts in the data, also behaved noticeably less well than the other procedures for these simulations. A perhaps more reasonable speed-up is based on the observation that for large changes that can be detected with small backgrounds, we want to estimate the background as accurately as possible. But for relatively large backgrounds used to detect relatively small changes, determining the exact background does not seem to be important. Suppose then, that in moving a candidate chang-point to the right, to determine k , we choose $k = j + 1$, then choose a new value of k recursively as the old value plus $\max(1, [(k - j)/10])$, where $[x]$ denotes the largest integer less than or equal to x . Thus, for $k - j < 20$, we choose every integer, then every second integer for $k - j < 30$, etc. The computational complexity of this procedure is of order $m(\log(m))^2\ell^2$, where $\ell = 10$.

Table 6 gives a brief summary of the modified procedure of the preceding paragraph in comparison with wild binary segmentation.

Table 5: Random Change-points, $m = 500$, $b_1 = 4.83$, $b_{Seq} = 4.33$, $b_{WBS} = 4.565$. Easy denotes the number of repetitions where all methods detected the correct number of change-points; Impossible gives the number of repetitions where no method detected the correct number of change-points.

Number	Min ($k - i$)	Seq	WBS	Easy/Impossible
0	479/0/22	481/0/29	465/0/37	455/8
3	421/68/18	420/52/45	408/63/38	377/50
5	337/184/11	371/140/16	319/174/38	294/111
8	201/450/2	249/337/25	211/412/21	161/212

Table 6: Random Change-points, $m = 500$, $b_1 = 4.83$, $b_{\text{WBS}} = 4.565$. Search algorithm modified as described in the text

Number	Min ($k - i$)	WBS
3	82/17/2	81/15/5
5	64/39/6	60/37/9
8	37/90/4	36/78/9

It is difficult to say that one of the procedures has definitively outperformed the others. The procedure based on (1) is the only one that in theory has control over the false positive rate. One could argue that it becomes overly conservative when the number of change-points is large, since the effective length of the region where false positives can occur is then relatively small. It is difficult to say on the basis of these simulations whether Seq or WBS commits more false positive errors; but it seems clear that WBS has less power. It might be interesting to calibrate these three with regard to their false positive rate. But the simulation is already time consuming, and anything we learn for the parameter choices of Table 6 may not hold in other scenarios.

5 Array CGH data

In this section we present examples involving change-points in copy number variation (CNV) from array CGH data.

We first consider the test cases GBM29 and GBM31 used by Lai *et al.* (2005), to compare different methods of segmentation, and subsequently by Du, Kao and Kou (2016).

For GBM29, the total length of the sequence is 193. The theoretical 0.05 thresholds for the procedure based on (1) and the pseudo-sequential procedure are 4.53 and 4.07, respectively. The estimated standard deviation is 0.76. Both algorithms detect the same set of change-points at

$$81, 85, 89, 96, 123, 133.$$

For GBM31, the length of the sequence is 797. The theoretical 0.05 thresholds are now 4.96 and 4.45, respectively. The estimated standard deviation is 0.38. Both algorithms again detect the same set of change-points at

$$317, 318, 538, 727, 728.$$

For BM31 the third change-point is a relatively small increase; the first two and last two are spikes that could be outliers. These two sets of detections are the same as those obtained by the method of Picard *et al.* (2005) and similar to those obtained by the method of Olshen *et al.* (2004). These methods were judged on the basis of simulations by Lai *et al.* (2005) to perform better than several other methods. Lai *et al.* (2005) also provide some additional discussion of the scientific meaning of these detections.

Du, Kao and Kou (2016) have used these data to test their Bayesian method. They detected essentially the same change-points in GBM29 but appear to have detected substantially more in GBM31, although the meaning of “detected” does not appear to be clearly specified.

We have also tested our methods on the BT474 cell line data from Snijders *et al.* (2003). See Pollack *et al.* (1999, 2002) for a different experimental technique involving BT474 and a discussion of the implications for breast cancer. This cell line has also been used by, e.g., Zhao *et al.* (2004), who based their experimental technique on SNPs rather than array CGH.

For a scan of the entire genome we detect 63 change-points with the statistic (1) at a 0.05 genome wide significance threshold of $b = 5.2$; and we detect 67 using the pseudo-sequential procedure. Since the estimated standard deviation of the observations can vary considerably from one chromosome to the next, we consider also several individual chromosomes.

Particularly interesting are chromosome 17, where an amplification at 17q23 appears to have implications for breast cancer, and chromosome 20, which contains a complex sequence of CNV. For chromosome 17, there are $m = 87$ observations, the standard deviation is 0.51; the 0.05 detection threshold is approximately 4.22. There is an increase in copy number at the 36th observation (17q11.2-12), with a change back to baseline just two observations later. There is a second increase at the 51st observation (17q21.3) and a return to the baseline at the 67th (17q23). Chromosome 20 contains $m = 85$ observations, and the standard deviation is 0.59. The threshold is again 4.22. The statistic (1) detects a *decrease* in copy number from the 39th (20q11.2) to the 52nd observation. This is followed by a substantial increase from the 53rd (20q13) to the 69th (20q13.1), except for a single outlier at the 58th. From the 69th observation there is an even larger increase until the 82nd (20q13.3), then a return to roughly the baseline value for the last three observations. The statistic (1) detects only a single change at the 58th observation, where there is a single much smaller observation in an otherwise large amplification. Also interesting are chromosomes 4, 5 and 11, all of which have several changes, and some of the changes are followed by a second change after only a few observations. On chromosome 4 there

are 162 observations, and at a 0.05 significance level (1) detected changes at 7, 8, 59, 61, 141, 143, 155. On chromosome 5 there were 99 observations, with changes detected at 25, 45, 51, 54, 65, 85, and 91. On chromosome 11 there are 184 observations and changes detected at 6, 90, 124, 139, 144, 162, 163, 164, 165.

Remark. In studying copy number variation it is customary to plot the locus by locus measurements, which should be about equal to zero when the copy number is two, with positive values indicative of amplifications and negative values indicative of deletions. There are advantages to plotting the consecutive partial sums and looking for a change in slope to indicate an increase or decrease in copy number. This plot is substantially smoother, and changes in slope that are candidates for change-points in copy number are easy to see. The disadvantage is that it is sometimes difficult to infer the sections of normal copy number, which are regions where the slope is near zero although it is never exactly equal to zero.

6 Exponential Families

In this section, we assume X_1, \dots, X_m are independent and from a one-parameter exponential family of distributions $\{F_\theta : \theta \in \Theta\}$ where

$$\frac{dF_\theta}{d\mu}(x) = \exp(\theta x - \psi(\theta)), \quad x \in \mathbb{R}, \quad \theta \in \Theta,$$

μ is a σ -finite measure on the real line and Θ is an open interval. For $0 \leq i < j < k \leq m$, the likelihood ratio statistic to test whether j is a change-point in the local background $(i, k]$ is

$$\begin{aligned} \ell_{i,j,k} = & (j-i) \sup_{\theta_1} \left(\theta_1 \frac{S_j - S_i}{j-i} - \psi(\theta_1) \right) + (k-j) \sup_{\theta_2} \left(\theta_2 \frac{S_k - S_j}{k-j} - \psi(\theta_2) \right) \\ & - (k-i) \sup_{\theta} \left(\theta \frac{S_k - S_i}{k-i} - \psi(\theta) \right). \end{aligned}$$

In the following, we use \mathbb{P}_θ (\mathbb{E}_θ resp.) to denote the probability (expectation resp.) calculated when $X_1 \cdots \sim F_\theta$. Following the proof of (4), we suggest

the following approximation to the p -value of $\max_{i,j,k} \ell_{i,j,k}$:

$$\begin{aligned} & \mathbb{P}_\theta\left(\max_{\substack{i < j < k \\ m_0 \leq j-i, k-j \leq m_1}} \ell_{i,j,k} \geq \frac{b^2}{2}\right) \\ & \sim \varphi(b) \sum_{n_1=m_0}^{m_1} \sum_{n_2=m_0}^{m_1} (m - n_1 - n_2) \sum_{\theta_1, \theta_2} \frac{a(\theta_1, \theta) a(\theta_1, \theta_2) a(\theta, \theta_2)}{[n_1(\theta_1 - \theta)^2 \psi''(\theta_1) + n_2(\theta_2 - \theta)^2 \psi''(\theta_2)]^{1/2}} \end{aligned} \quad (8)$$

where the third summation is over two pairs of $\theta_1 < \theta_2$, which are assumed to exist (see the remark below), solving

$$\begin{cases} \psi'(\theta_1)n_1 + \psi'(\theta_2)n_2 = \psi'(\theta)(n_1 + n_2), \\ n_1[\theta_1\psi'(\theta_1) - \psi(\theta_1)] + n_2[\theta_2\psi'(\theta_2) - \psi(\theta_2)] - (n_1 + n_2)[\theta\psi'(\theta) - \psi(\theta)] = b^2/2, \end{cases} \quad (9)$$

and for $\theta_1 < \theta_2$,

$$a(\theta_1, \theta_2) = \exp\left(-\sum_1^\infty n^{-1} E_{\theta_2} e^{-[(\theta_2 - \theta_1)S_n - n(\psi(\theta_2) - \psi(\theta_1))]^+}\right).$$

We use Theorem 8.51 of Siegmund (1985) and Theorem A of Tu and Siegmund (1998) to compute $a(\theta_1, \theta_2)$ numerically for nonarithmetic and arithmetic random variables respectively.

Remark. For those n_1 and n_2 such that the solutions to (9) do not exist, we first find the smallest $\theta' > \theta$ such that the solutions to (9) with θ replaced by θ' exist. We denote the solutions by θ'_1 and θ'_2 . Then the proposed approximation is the RHS(8) with $\theta, \theta_1, \theta_2$ replaced by $\theta', \theta'_1, \theta'_2$ respectively, and multiplied by $\mathbb{P}_\theta(S_{n_1+n_2}/(n_1 + n_2) \geq \psi'(\theta'))$.

6.1 Simulations

We first consider the exponential distribution with rate λ . Observing that in (8), both the probability and its approximation do not depend on λ , we choose $\lambda = 1$ without loss of generality. We fix $m_0 = 1$. In Table 7, with different values of m, m_1 and b , p denotes the RHS(8) and \hat{p} denotes the simulated p -value with 2000 repetitions. We see from Table 7 that our approximation to the p -values are reasonably accurate, especially when m and m_1 are large. A normal approximation as given above would also be quite reasonable, especially for larger m_1 and m . For example, for the last line of Table 7 our normal approximation gives the probability 0.053.

Table 7: Exponential distribution with rate λ .

λ	m	m_1	b	p	\hat{p}
1	500	50	4.72	0.049	0.061
1	500	100	4.78	0.048	0.053
1	1000	100	4.95	0.047	0.0475

Next, we consider the inverse Gaussian distribution with fixed shape parameter $\lambda = 10$. We fix $m_0 = 1$. With different values of the mean μ , m, m_1 and b , p denotes the RHS(8) and \hat{p} denotes the simulated p -value with 2000 repetitions. We can see from Table 8 that both the theoretical and simulated p -values are reasonably robust against the mean μ .

Table 8: Inverse Gaussian distribution with shape parameter $\lambda = 10$.

μ	m	m_1	b	p	\hat{p}
1	300	30	4.5	0.059	0.053
5	300	30	4.5	0.041	0.0495
1	500	100	4.78	0.049	0.037
5	500	100	4.78	0.035	0.0305
1	1000	100	4.95	0.049	0.0495
5	1000	100	4.95	0.036	0.0335

We have also conducted simulations for Bernoulli observations. For the detection of CG rich regions mentioned in the introduction, the sequences are very long and the exact boundary between regions has little biological significance. Hence one often forms groups of consecutive Bernoulli variables. Following Ehaik, Graur and Josić (2010), we have used groups of 33 consecutive Bernoulli variables. Since the values of the Bernoulli parameters p are usually neither extremely small or extremely large (cases that might indicate a Poisson approximation), we have tentatively assumed that we can use the theory developed above for the normal distribution. Since the Bernoulli variances must be estimated locally in each homogeneous region, it turns out that the skewness of the binomial distribution when p is not in the immediate neighborhood of $1/2$ can make an approximation of the distribution of the scaled value of $[S_j - S_i - (j - i)(S_k - S_i)/(k - i)]$ by a normal distribution unsatisfactory, unless the size of the groups is rela-

tively large. Consequently we have used the signed square roots of the log likelihood ratio statistics, which behave very much like a Gaussian process. Since simulations of this process indicate a satisfactory approximation of the significance level and no new insights into power, we omit the details.

We can deal with Poisson distributed observations similarly, and again simulations support an approximation based on a normal approximation of the signed square root of the log likelihood ratio statistic. For 500 observations, $b = 4.83$, and the mean of the Poisson distribution equal to 10, 400 simulations gave the significance value 0.0475, when our normal approximation gives the value 0.05. Calculation of Kullback-Leibler information suggests that for detecting changes from 10 to 20 and back to 10 in well separated intervals, lengths of 6 and 7 are borderline detectable. Several simulations of this case involving two pairs of change-points lead to successful detections of all four change-points, while the differences between the estimates and the true values totaled 1-3 observations.

7 Confidence Regions

We now return to our basic assumption of independent normal observations with mean values forming a step function and variance equal to one. For a given value of M , we can use the likelihood ratio statistic to construct a joint confidence region for the change-points $\tau = (\tau_1, \dots, \tau_M)$ and mean values $\mu = (\mu_1, \dots, \mu_{M+1})$. For testing a putative value of the positions of change points and the corresponding mean values, the maximum log likelihood ratio statistic is

$$T_{\tau, \mu} = \max_{t_1, \dots, t_M} \sum_{k=1}^{M+1} \frac{(S_{t_k} - S_{t_{k-1}})^2}{2(t_k - t_{k-1})} - \sum_{k=1}^{M+1} \left[\mu_k (S_{\tau_k} - S_{\tau_{k-1}}) - \frac{\mu_k^2}{2} (\tau_k - \tau_{k-1}) \right], \quad (10)$$

where the maximum is taken over $0 < t_1 < \dots < t_M < m$ and $S_i = \sum_{j=1}^i X_j$ for $0 \leq i \leq m$. The $1 - \alpha$ confidence region consists of those τ and μ such that $T_{\tau, \mu} \leq b_{\tau, \mu}$ where

$$\mathbb{P}_{\tau, \mu}(T_{\tau, \mu} > b_{\tau, \mu}) = \alpha. \quad (11)$$

If we are only interested in the confidence region of τ and treat μ as nuisance parameter, the maximum log likelihood ratio statistic is

$$T_{\tau} = \max_{t_1, \dots, t_M} \sum_{k=1}^{M+1} \frac{(S_{t_k} - S_{t_{k-1}})^2}{2(t_k - t_{k-1})} - \sum_{k=1}^{M+1} \frac{(S_{\tau_k} - S_{\tau_{k-1}})^2}{2(\tau_k - \tau_{k-1})}. \quad (12)$$

See Section 4 of Siegmund (1988) for the corresponding statistic for exponential families in the case of one change point. By sufficiency the conditional distribution of T_τ given $\{S_{\tau_k} : 1 \leq k \leq M+1\}$ does not depend on θ . Therefore, a $1 - \alpha$ confidence set for the change points is the set of τ such that $T_\tau \leq b_{\tau, s_{\tau_1}, \dots, s_{\tau_M}}$ where

$$\mathbb{P}_\tau(T_\tau > b_{\tau, s_{\tau_1}, \dots, s_{\tau_M}} | \tau, S_{\tau_1} = s_{\tau_1}, \dots, S_{\tau_M} = s_{\tau_M}) = \alpha. \quad (13)$$

In the case there is only one change point, i.e., $M = 1$, the exact value of the left-hand side of (13) was given by Worsley (1986) for exponential random variables. Asymptotic approximations for the left-hand side of both (11) and (13) were given by Siegmund (1988) for exponential families. Since the asymptotic approximations in Siegmund (1988) seem difficult to generalize to the case where $M \geq 2$, here we use a different approach to obtain asymptotic approximations for the left-hand side of both (11) and (13) for $M \geq 1$.

In the case $M = 1$, Siegmund (1988) showed that the expected size of the confidence interval for τ constructed by the likelihood ratio statistic is asymptotically half as large as that constructed from the distribution of the maximal likelihood estimator of τ . He also showed by simulation that the expected size of the likelihood ratio confidence interval is close to those constructed by Bayesian approaches. We will compare our confidence region with those constructed by the approach of Frick, Munk and Sieling (2014).

7.1 Tail approximations

To construct the joint confidence region for the change points and the corresponding parameters, for each τ and μ , we need to find $b_{\tau, \mu}$ such that $\mathbb{P}_{\tau, \mu}(T_{\tau, \theta} > b_{\tau, \mu}) = \alpha$ where $T_{\tau, \theta}$ is defined in (10). The following theorem gives an approximation for large b . We suppose that the putative change-points are close enough to the true change-points so that the maximum can be taken over a relatively small neighborhood ($|t_k - \tau_k| \leq n_k$) of the putative change-points. We also require that the change-points to be sufficiently well separated from one another.

Theorem 1. Let $\tau = \{\tau_1, \dots, \tau_M\}$ and $\mu = \{\mu_1, \dots, \mu_{M+1}\}$ be defined as above; and let $T_{\tau, \mu}$ be defined as in (10). Define $\delta_k = \mu_{k+1} - \mu_k$ for $1 \leq k \leq M$ and $m_k = \tau_k - \tau_{k-1}$ for $1 \leq k \leq M+1$. Suppose that $|\delta_k|$ is bounded away from 0 and ∞ . Suppose further that

$$1 \ll b \ll n_k \ll (m_k \wedge m_{k+1})/b. \quad (14)$$

Let $U_{t,\mu}$ denote the expression on the RHS(10) before taking the maximum. We have,

$$\mathbb{P}_{\tau,\mu}(\max_{t:|t_k-\tau_k|\leq n_k} U_{t,\mu} > b) \sim \mathbb{P}(\sum_{k=1}^M W_k + \frac{1}{2}\chi_{M+1}^2 > b), \quad (15)$$

where $W_1, \dots, W_M, \chi_{M+1}^2$ are independent, χ_{M+1}^2 is a chi-squared random variable with $M+1$ degrees of freedom, and W_k satisfies

$$\mathbb{P}(W_k \geq x) = 2\nu(|\delta_k|)e^{-x} - \nu^2(|\delta_k|)e^{-2x}$$

for $1 \leq k \leq M$.

We have a similar approximation for the left-hand side of (13).

Theorem 2. Let T_τ be defined as in (12) with the maximum taken over $|t_k - \tau_k| \leq n_k$ for $1 \leq k \leq M$. Define $\hat{\delta}_k = \hat{\mu}_{k+1} - \hat{\mu}_k$ for $1 \leq k \leq M$, $\hat{\mu}_k = (S_{\tau_k} - S_{\tau_{k-1}})/(\tau_k - \tau_{k-1})$ and $m_k = \tau_k - \tau_{k-1}$ for $1 \leq k \leq M+1$. Suppose that $|\hat{\delta}_k|$ is bounded away from 0 and ∞ . Suppose further that

$$1 \ll b \ll n_k \ll (m_k \wedge m_{k+1})/b.$$

We have

$$\mathbb{P}_\tau(T_\tau > b | S_{\tau_1} = s_{\tau_1}, \dots, S_{\tau_M} = s_{\tau_M}) \sim \mathbb{P}(\sum_{k=1}^M W_k > b)$$

where W_1, \dots, W_M are independent have the same distributions as in Theorem 1.

It is easy to evaluate the distributions of $\sum W_k$ and $\sum W_k + \chi_{M+1}^2$ by Fourier inversion, for values of M up to about 100, and by exponential tilting if still larger values of M are of interest. For simplicity assume that $\delta_k = \delta$ for all k . Let $\nu = \nu(\delta)$ and $\hat{f}(\lambda) = (1 - \nu)^2 + 2\nu/(1 + \sqrt{-1}\lambda) - 2\nu^2/(2 - \sqrt{-1}\lambda)$ denote the characteristic function of W_k . Let $\hat{g}(\lambda)$ be the characteristic function of a χ_{M+1}^2 random variable. Finally, let $h(\lambda) = \hat{f}^m(\lambda) * \hat{g}(\lambda)[1 - \exp(\sqrt{-1}\lambda b)]/(1 + \sqrt{-1}\lambda)$. Then the probability on the right hand side of (15) equals $1 - \int_0^\infty \mathbf{Re}[h(\lambda)]d\lambda/\pi$.

7.2 Simulations

Here we use simulations to check the accuracy of the approximation (15). We fix the number of change-points to be $M = 2$. For different values of δ_1 and δ_2 , we compute the threshold b such that RHS(15) equals 0.05.

In the simulations, we fix the length of the sequence to be $m = 210$ and the locations of change-points to be $\tau_1 = 70, \tau_2 = 140$, and we let the mean values be $\mu_1 = 0, \mu_2 = \delta_1, \mu_3 = \delta_1 + \delta_2$. In Table 9, \hat{p} denotes the probability on the LHS(15) with $n_k = m$ and is based on 10000 repetitions each. From Table 9, we can see that the approximation (15) is reasonably accurate for the range $1 < |\delta| < 2$. Smaller $|\delta|$, requires larger m to make the approximation more accurate.

Table 9: Confidence Region Probabilities: $M = 2, \tau_1 = 70, \tau_2 = 140, m = 210$

δ_1	δ_2	b	\hat{p}
0.75	0.75(-0.75)	7.58	0.0536(0.0567)
1	1(-1)	7.22	0.0447(0.0498)
1.25	1.25(-1.25)	6.91	0.0472(0.0457)
1.5	1.5(-1.5)	6.62	0.0457(0.0468)
2	2(-2)	6.14	0.0419(0.0471)
1.5	.75(-.75)	7.07	0.0475(0.0495)

In Table 10 we have selected some parameter values from Table 3. The upper part of the table, like Table 9, gives the estimated coverage probability based on 10000 simulations for examples where the threshold b has been selected so our theoretical approximation gives the probability 0.05. The lower part of the table gives the probability from 1000 simulations that the indicated values of t_1, t_2 are *not* contained in the confidence region. We have chosen values of t_i for which this probability is about 0.5, so one can regard the difference between t_i and τ_i as a rough measure of the size of the confidence region when all other parameters are set to their correct values.

The rows beginning with 0.65 are particularly interesting, since they show that the relatively small change at $\tau_1 = 138$ compared with very large change at $\tau_2 = 225$ leads to substantially more uncertainty in the value of τ_1 compared to the value of τ_2 .

7.3 Comparison with other confidence intervals

Frick, Munk and Sieling (2014) suggested a different method to construct a confidence region jointly for the change-points and the mean values of the observations in the segments connecting those change-points. For each candidate set of change-points τ and mean values μ , they suggest an application

Table 10: Likelihood ratio based joint confidence intervals. \hat{p} is the simulated probability that the parameters t_1 and t_2 are rejected when the true parameter values are τ_1 and τ_2 . Nominal confidence level is 0.05. Simulations are based on 10000 (1000) repetitions in the first four (last 12) rows.

δ_1	δ_2	b	τ_1, τ_2	t_1, t_2	\hat{p}
2.13	1.33	6.4	9, 33	9, 33	0.049
2.5	4.0	5.35	87, 104	87, 104	0.051
0.65	2.5	6.65	138, 225	138, 225	0.047
1.73	2.13	6.23	57, 66	57, 66	0.049
2.13	1.33	6.4	9, 33	7, 33	0.59
2.13	1.33	6.4	9, 33	11, 33	0.58
2.13	1.33	6.4	9, 33	9, 29	0.47
2.13	1.33	6.4	9, 33	9, 37	0.44
0.65	2.5	6.65	138, 225	138, 227	0.75
0.65	2.5	6.65	138, 225	138, 223	0.73
0.65	2.5	6.65	138, 225	120, 225	0.49
0.65	2.5	6.65	138, 225	156, 225	0.46
2.5	4.0	5.35	87, 104	87, 102	0.43
2.5	4.0	5.35	87, 104	87, 106	0.44
2.5	4.0	5.35	87, 104	86, 104	0.89
2.5	4.0	5.35	87, 104	88, 104	0.89

of their multiscale statistic

$$\max \left(\frac{|S_j - S_i - (j - i)\mu|}{(j - i)^{1/2}} - [2 \log(3m/(j - i))]^{1/2} \right) \quad (16)$$

where the maximum is taken over all $i < j$ within one of the segments of $(0, \tau_1], \dots, (\tau_M, m]$, μ is the corresponding mean value in the segment. This is in effect a test of the hypothesis that there are no additional change-points in the segments $(0, \tau_1], \dots, (\tau_M, m]$. Worsley (1986) discusses a similar idea under the assumption that there is a single change-point, and one is interested only in a confidence region for the change-point, not a joint confidence region for change-points and means.

Here we compare our confidence region defined by (11) with that using (16). We set $m = 200$, $\tau_1 = 50$, $\tau_2 = 100$ and consider values of the δ_i that are large enough that most of the time we will see correctly that there are two change-points. The problem becomes one of locating them

and estimating the mean values. For our confidence regions, we choose the thresholds $b_1 = 7.2$ so that the probability in (15) equals 0.05, which is confirmed by simulation. Moreover, for the statistic (16), we choose the threshold $b_2 = 1.7$ for which a 20000 repetition simulation experiment gives the probability 0.05. This threshold is slightly larger than the theoretical approximation provided by Fang and Siegmund (2015).

Since a direct comparison of these regions in terms of size is conceptually complicated and technically demanding, we use the relation of confidence regions to hypothesis testing to compare them in terms of power. Under specific hypothetical, but incorrect, values of the change-points and mean values the power of the test of the true values represents the probability that the hypothetical values do not lie in the confidence region. Hence the procedure with larger power is preferred. It seems clear that for the parameter settings analysed, the likelihood ratio procedure is preferable.

Table 11: Power to Detect Departure from True Parameter Values: $\tau = (50, 100)$ and μ as given; t and ξ are hypothesized values of τ and μ . The subscript 1 indicates the likelihood ratio procedure, while 2 indicates the procedure based on (16).

μ	ξ	t	Power ₁	Power ₂
0.0,1.0,0.0	0.0,1.0,0.0	55,95	0.64	0.14
0.0,1.0,0.0	0.1, 0.9, -0.2	55,95,	0.87	0.57
0.0,1.0,0.0	0.1, 0.9, -0.2	40, 100	0.75	0.48
0.0,1.2,2.0	0.0,1.2,2.0	47,105	0.47	0.085
0.0,1.2,2.0	0.0,1.5,1.9	47,105	0.75	0.52
0.0,1.5,0.75	0.1,1.4,0.9	40,97	0.96	0.77
0.0,1.5,0.75	0.0,1.5,0.75	44,98	0.81	0.39
0.0,1.2,-.1	0.1,1.1,0.1	48,103	0.68	0.36
0.0,1.1,0.1	-.2,1.0,0.0	52,115	0.91	0.56
0.0,1.0,2.0	-0.1,1.1,2.1	45,110	0.87	0.33

7.4 A Different Approach to Estimation.

With our approach of determining suitable local backgrounds for the detection of change-points, it is natural to consider those backgrounds as interval estimates for the change-points. In this section we report the results of some simulations that use the interval (i, k) , selected as the smallest value of $k - i$

for which $\max_j |Z_{i,j,k}|$ exceeds a suitable threshold, as an interval estimator of the change-point j . In principle we could also use the background associated with the largest value of $|Z|$, but our numerical experiments show that this leads to much larger intervals with only a slight gain in accuracy.

A few selected examples are given in Table 12. Note that the change-points at 138 and at 299 are the most difficult to detect (Table 3). When detected, the change-point at 138 can be poorly estimated and has a very long background interval. The difficulty with detecting the change-point at 299 is that it is very close to the change-point at 308, so the associated value of $k - j$ is necessarily small, while the value of $j - i$ can be fairly large. Although the change-point at 232 is easily detected, it is somewhat isolated from the nearest change-point at 308, and hence its background interval can be reasonably large.

We also simulated pairs of examples, starting with the same Monte Carlo seed. The first of the pair had the same change-points and values of δ as in Table 14. For the second, the first four values of delta were all positive; the last two negative. For detection based on the shortest local background, the detected change-points and their backgrounds were often very similar. These results give additional credence to our theoretical analysis, which was essentially limited to consideration of one change-point at a time against a suitable local background.

A summary of a more extensive simulation is given in Table 13. Following Olshen *et al.* (2004), we have added a “local trend,” which in array based CNV analysis arises as an artifact of local features of the DNA. For this simulation the local trend is $0.25a_0 \sin(a_1\pi n)$, which is added to the mean of the n th observation.

The averages reported in Table 13 are consistent with the individual examples of Table 12. The local trends can have a substantial effect on the power to detect a change-point, but relatively little effect on the conditional expected length of the background interval. Although the intervals obtained in this analysis appear to be substantially longer than those based on our confidence intervals, they also appear to contain the true change-points substantially more frequently.

8 Discussion

The segmentation procedures we have studied in this paper are local thresholding procedures. For subsets of intervals (i, k) a statistic $Z_{i,j,k}$ for $i < j < k$ is studied to see if there is evidence of a change-point at j , by comparing

Table 12: Examples of Shortest Backgrounds as Interval Estimators: 138,225,242,299,308,332 with changes in the mean value: 0.65,2.5, -4.0, 1.73, -2.13, 1.33. Threshold is $b = 4.83$.

i	j	k	$k - i$
79	138	173	94
216	225	232	16
239	242	244	5
289	298	307	18
318	334	350	32
72	151	167	95
215	225	233	18
240	242	244	4
307	331	415	108
214	225	232	28
236	242	244	8
283	299	305	22
299	307	311	12
323	331	346	21
48	114	165	117
218	225	231	13
238	242	245	7
271	299	306	35
299	307	321	22
315	331	345	20
69	139	169	100
219	225	230	11
238	242	245	7
300	307	310	10
307	325	341	34
78	141	207	129
221	225	238	17
239	242	244	5
294	299	302	9
300	307	309	9
315	336	357	43

Table 13: Averages of Shortest Backgrounds as Interval Estimators: 138,225,242,299,308,332 with changes in the mean value: 0.65,2.5, -4.0, 1.73, -2.13, 1.33. Local trend parameters a_0, a_1 as explained in the text. Threshold is $b = 4.83$.

$a_0 = a_1 = 0$	Power	0.73	1.00	1.00	0.68	0.84	0.92
	Conditional Exp Length	115	13	5.9	25	16	44
$a_0 = 0.1, a_1 = 0.01$	Power	0.62	1.00	1.00	0.65	0.84	0.93
	Conditional Exp Length	115	13	5.9	25	16	46
$a_0 = 0.1, a_1 = 0.025$	Power	0.73	1.00	1.00	0.62	0.81	0.93
	Conditional Exp Length	115	13	5.9	24	16	42
$a_0 = 0.2, a_1 = 0.01$	Power	0.52	1.00	1.00	0.60	0.87	0.93
	Conditional Exp Length	112	13	5.9	24	16	48
$a_0 = 0.2, a_1 = 0.025$	Power	0.74	1.00	1.00	0.56	0.78	0.94
	Conditional Exp Length	113	13	6.0	23	16	39

values of the statistic with a threshold designed to control the probability of a false positive error. Our pseudo-sequential procedure, leaves i fixed at 0 or at the most recently discovered candidate change-point, then sequentially with respect to k examines $\max_{i < j < k} Z_{i,j,k}$ until it exceeds a suitable threshold. The statistic based on (1) has better false positive control, although it requires a relatively large threshold, and hence loses some power compared to the pseudo-sequential version.

We have compared our suggested procedures to the classical binary segmentation procedure, the Wild Binary Segmentation procedure of Fryzlewicz (2014), and the SaRa procedure of Niu and Zhang (2012). Of these, the first seems competitive with our procedures, although the false positive error control is not well understood. The second performs well when the change-points are well separated, but poorly when they are close together. We also made less systematic comparisons to the multiscale procedure of Frick, Munk and Sieling (2014) and the Bayesian procedure of Du, Kao and Kou (2016), which with default parameters seemed uncompetitive so were not included in our detailed simulations.

These procedures are all local, in the sense that they examine subsets of the data for evidence of change-points within the subset. They are relatively easily evaluated in comparison to global procedures that consider all possible partitions of the data, e.g., the modified BIC procedure of

(Zhang and Siegmund (2007)). This has led some authors, for example, Fryzlewicz (2014) and Niu and Zhang (2012) to consider two stage procedures. A local procedure, presumably with a fairly low threshold, is used to suggest candidate change-points, and is followed by a global procedure focused on the candidates to determine the change-points to include in the final model. The local procedures we have introduced can also be used as part of such a two-stage procedure. It would also be possible to use the pseudo-sequential procedure, which is relatively fast but has less adequate false positive control, to select a candidate set of change-points, presumably at a fairly low threshold, and then use the slower procedure that considers trios of indices i, j, k to eliminate some of the candidates and improve the accuracy of others. A speed-up of the procedure based on $\max_{i < j < k} |Z_{i,j,k}|$ is described in the text.

We have used the likelihood ratio statistic to obtain approximate confidence regions for the locations of the change-points or jointly for the mean values and the locations. We also investigated by simulation the use of the local backgrounds of the detection procedure based on (1) to obtain an informal assessment of the simultaneous accuracy of the detected change-points.

In a related, but somewhat different problem from the one studied in this paper, change-points are expected to come in pairs where paired changes in the mean have roughly equal magnitude and opposite signs. An interesting application is detection of inherited copy number variations, which involve short genomic segments where either a gain or a loss in copy number is followed by a subsequent return to the baseline value. See, for example, Zhang *et al.* (2010). Problems having this structure also arise in multidimensional time, e.g., brain imaging, where one wants to detect a local change of level in a random field. The procedures of this paper can be used to detect paired changes, but it should be possible to do better by taking account of the special structure of the anticipated signal.

Two thresholding procedures that are designed for detection of such paired change-points but have also been used for segmentation of an arbitrary pattern of change-points are the circular binary segmentation procedure (CBS) of Olshen *et al.* (2004) and the multiscale procedure of Frick, Munk and Sieling (2014). The procedures use the maximum over values of $j \in [i, k]$ and $1 \leq n \leq k - j$ of the statistic

$$|S_{j+n} - S_j - n(S_k - S_i)/(k - i)|/[n(1 - n/(k - i))]^{1/2} - r[2 \log(3m/n)]^{1/2},$$

where the background $0 \leq i < k \leq m$ is selected in advance, starting with initial values $i = 0, k = m$. The value $r = 0$ gives the CBS statistic,

while $r = 1$ gives the multiscale statistic of Frick, Munk and Sieling (2014). (In principle, one could also consider these statistics with locally adaptive backgrounds, although they are computationally onerous.) Some limited numerical experimentation indicates that when applied iteratively with a fixed threshold, these procedures are often very effective. In some cases, in an interval (i, k) the maximum over j and n may be achieved by a value $j \approx i$ or by $j + n \approx k$. This can indicate that only one change-point has been detected on that iteration, but some judgment may be required to reach this conclusion. When there are multiple change-points with the mean value increasing (or decreasing) at consecutive change-points, these statistics seem somewhat prone to false positive errors. We expect to study them and related procedures in a future paper.

We also considered briefly a smoothed version of the Niu-Zhang statistic. This has some theoretical advantages and does not seem to have substantially less power than the unsmoothed version. Smoothing can be particularly useful for detecting local signals in a multidimensional random field, since smooth random fields seem more tractable for dealing with the large set of possible signal shapes that can occur in several dimensions (e.g., Adler and Taylor (2007)).

Appendix A

In this appendix, we prove (4) under the conditions that

$$b \rightarrow \infty, \quad m_0, m_1 - m_0 \asymp b^2, \quad b^2 \ll m \ll \frac{e^{b^2/2}}{b^3}.$$

The derivation of (5) and (6) are similar, therefore omitted. The claims stated in the proof will be proved below.

For $m_0 \leq h_1, h_2 \leq m_1$, define

$$d_1 = d_1(h_1, h_2) = \frac{b}{2h_1 \sqrt{\frac{1}{h_1} + \frac{1}{h_2}}},$$

$$d_2 = d_2(h_1, h_2) = \frac{b}{2} \sqrt{\frac{1}{h_1} + \frac{1}{h_2}},$$

and

$$d_3 = d_3(h_1, h_2) = \frac{b}{2h_2 \sqrt{\frac{1}{h_1} + \frac{1}{h_2}}}.$$

Since $h_1, h_2 \asymp b^2$, we have $d_i, \nu(2d_i) \asymp 1$ for $i = 1, 2, 3$. Therefore, denoting the right-hand side of (4) by p , we have

$$p \asymp mm_1^2 \frac{\varphi(b)}{b}, \quad (17)$$

which converges to zero under the conditions $m_1 \asymp b^2$ and $m \ll e^{b^2/2}/b^3$.

Claim 1. The two-sided probability on the left-hand side of (4) is asymptotically equal to twice the one-sided probability

$$P(|Z_{ijk}| \geq b) \sim 2P(Z_{ijk} \geq b).$$

Following an argument in Zhang and Liu (2011), we write

$$\begin{aligned} & P\left(\max_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} Z_{ijk} \geq b\right) \\ &= \sum_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} P(Z_{ijk} \geq b, Z_{ijk} = \max_{\substack{0 \leq r < s < t \leq m \\ m_0 \leq s-r, t-s \leq m_1}} Z_{rst}) \\ &= \sum_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} \int_b^\infty P\left(\max_{\substack{0 \leq r < s < t \leq m \\ m_0 \leq s-r, t-s \leq m_1}} Z_{rst} \leq b+x \mid Z_{ijk} = b+x\right) dP(Z_{ijk} = b+x) \\ &= \sum_{\substack{C \log b \leq i < j < k \leq m - C \log b \\ m_0 \leq j-i, k-j \leq m_1}} \int_b^{b+1} P\left(\max_{\substack{0 \leq r < s < t \leq m \\ m_0 \leq s-r, t-s \leq m_1}} Z_{rst} \leq b+x \mid Z_{ijk} = b+x\right) dP(Z_{ijk} = b+x) + R \end{aligned}$$

where C is a positive constant to be chosen in Claim 2 and the remainder R is of smaller order than p .

To prove that

$$P\left(\max_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} Z_{ijk} \geq b\right) \sim p/2,$$

we only need to show that for any $C \log b \leq i < j < k \leq m - C \log b$ such that $m_0 \leq j-i, k-j \leq m_1$, we have

$$P\left(\max_{\substack{0 \leq r < s < t \leq m \\ m_0 \leq s-r, t-s \leq m_1}} Z_{rst} \leq b \mid Z_{ijk} = b\right) \sim \prod_{i=1}^3 (2d_i^2) \nu(2d_i), \quad (18)$$

where d_i is defined above with $h_1 = j-i, h_2 = k-j$.

In the following we fix any i, j, k such that $C \log b \leq i < j < k \leq m - C \log b$ and $m_0 \leq j-i, k-j \leq m_1$ and prove (18). Let $h_1 = j-i, h_2 = k-j$.

Claim 2. There exists a large enough constant C such that

$$P\left(\max_{\substack{0 \leq r < s < t \leq m \\ m_0 \leq s-r, t-s \leq m_1, (|r-i| \vee |s-j| \vee |t-k|) \geq C \log b}} Z_{rst} > b | Z_{ijk} = b\right) = o(1).$$

From Claim 2, the maximum in (18) can be restricted to those r, s, t such that $|r-i|, |s-j|, |t-k| \leq C \log b$.

Next, we note that given $Z_{ijk} = b$, we have

$$\frac{S_j - S_i}{j - i} = \mu - 2d_1(1 + o(1)) \text{ and } \frac{S_k - S_j}{k - j} = \mu + 2d_3(1 + o(1)). \quad (19)$$

This, together with Theorem 1.6 of Diaconis and Freedman (1988) and the fact that $\log b \ll m_0$, implies that given $Z_{ijk} = b$,

$$X_{i+1}, \dots, X_{i+C \log b}, X_{j-C \log b+1}, \dots, X_j, X_{j+1}, \dots, X_{j+C \log b}, X_{k-C \log b+1}, \dots, X_k$$

are asymptotically mutually independent Gaussian variables with variance 1, the first half of the X 's have mean $\mu - 2d_1$ and the second half of the X 's have mean $\mu + 2d_3$. Let us first consider the case $r = i, s = j$ and $k < t \leq k + C \log b$ in (18). Note that $Z_{ijk} = b$ and $Z_{ijt} \leq b$ are equivalent to

$$h_1(S_k - S_j) - h_2(S_j - S_i) = bh_1h_2\sqrt{\frac{1}{h_1} + \frac{1}{h_2}} \quad (20)$$

and

$$h_1(S_t - S_k + S_k - S_j) - (h_2 + t - k)(S_j - S_i) \leq bh_1(h_2 + t - k)\sqrt{\frac{1}{h_1} + \frac{1}{h_2 + t - k}}. \quad (21)$$

Subtracting (20) from (21) and using Taylor's expansion and (19), we have that given $Z_{ijk} = b$, $Z_{ijt} \leq b$ is equivalent to

$$\sum_{l=1}^{t-k} [X_{k+l} - \mu - d_3(1 + o(1))] \leq 0.$$

Therefore,

$$P\left(\max_{k < t \leq k + C \log b} Z_{ijt} \leq b | Z_{ijk} = b\right) \\ P\left(\max_{1 \leq l \leq C \log b} (X_{k+l} - \mu - d_3(1 + o(1))) \leq 0\right).$$

Note that $X_{k+l} : l \geq 1$ are i.i.d. $\sim N(\mu, 1)$. Using the union bound, we have

$$P\left(\max_{l > C \log b} (X_{k+l} - \mu - d_3(1 + o(1))) > 0\right) = o(1).$$

Therefore,

$$\begin{aligned}
& P\left(\max_{1 \leq l \leq C \log b} (X_{k+l} - \mu - d_3(1 + o(1))) \leq 0\right) \\
& \approx P\left(\max_{l \geq 1} (X_{k+l} - \mu - d_3(1 + o(1))) \leq 0\right) \\
& \approx \sqrt{2} d_3 \nu^{1/2}(2d_3),
\end{aligned}$$

where the last equation is by Corollary 8.44 of Siegmund (1985).

Similar arguments for the other cases show that given $Z_{ijk} = b$, the event

$$\max_{\substack{0 \leq r < s < t \leq m \\ |r-i|, |s-j|, |r-k| \leq C \log b}} Z_{rst} \leq b$$

is asymptotically the same as the event that six random walks starting from 0 and stays below 0 before time $C \log b$. These random walks are asymptotically independent and have independent Gaussian increments with variance 1 and means $-d_1, -d_1, -d_2, -d_2, -d_3, -d_3$ respectively. Therefore,

$$P\left(\max_{\substack{0 \leq r < s < t \leq m \\ |r-i|, |s-j|, |r-k| \leq C \log b}} Z_{rst} \leq b \mid Z_{ijk} = b\right) \approx \prod_{i=1}^3 (2d_i^2) \nu(2d_i).$$

This proves (18).

Proof of Claim 1. We write

$$\begin{aligned}
& P\left(\max_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} |Z_{ijk}| \geq b\right) \\
& = P\left(\max_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} Z_{ijk} \geq b\right) + P\left(\max_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} \{-Z_{ijk}\} \geq b\right) \\
& \quad - P\left(\max_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} Z_{ijk} \geq b, \max_{\substack{0 \leq r < s < t \leq m \\ m_0 \leq s-r, t-s \leq m_1}} \{-Z_{rst}\} \geq b\right)
\end{aligned}$$

The first two terms are equal by symmetry. The third term is bounded by

$$\begin{aligned}
& \sum_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} P\left(Z_{ijk} \geq b, \max_{\substack{0 \leq r < s < t \leq m \\ m_0 \leq s-r, t-s \leq m_1}} \{-Z_{rst}\} \geq b\right) \\
& = \sum_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} P(Z_{ijk} \geq b) P\left(\max_{\substack{0 \leq r < s < t \leq m \\ m_0 \leq s-r, t-s \leq m_1}} \{-Z_{rst}\} \geq b \mid Z_{ijk} \geq b\right)
\end{aligned}$$

We first show that the conditional probability above tends to 0. To see this, we separate it into two terms according to whether $(r, t]$ overlaps with

$(i, k]$ or not. For the non-overlap terms, the conditional probability equals the unconditional probability, which goes to 0 as argued at the beginning of the proof of (4). For the overlap terms, we use the union bound. There are totally $O(b^6)$ overlap terms, each having a subgaussian probability in terms of b . Therefore, the conditional probability goes to 0 as well. By the above argument, the third term is of smaller order than

$$\sum_{\substack{0 \leq i < j < k \leq m \\ m_0 \leq j-i, k-j \leq m_1}} P(Z_{ijk} \geq b),$$

which is of the same order in (17) as the first two terms. Thus Claim 1 follows. \square

Proof of Claim 2. We separate the conditional probability into two terms according to whether $(r, t]$ overlaps with $(i, k]$ or not.

For the non-overlap terms, the conditional probability equals the unconditional probability, which goes to 0 as argued at the beginning of the proof of (4).

For the overlap terms, we use the union bound. There are totally $O(b^6)$ overlap terms. By choosing C to be large enough, each overlap term is can be bounded by $o(b^{-6})$.

This proves Claim 2. \square

Appendix B

In this appendix, we derive the approximation (7). There are mainly two approaches to derive approximations such as (7) for smooth random fields. One possibility uses a likelihood ratio transformation. been applied to a variety different problems, and has been For a systematic treatment and many examples see Yakir (2013) and the references cited there. An alternative approach would be to use the expected Euler characteristic (e.g., Adler and Taylor (2007)). Both approaches arrive at the following approximation for the probability in (7):

$$\mathbb{P}(\max_{0 < t < m, h_0 \leq h \leq h_1} |Z_{t,h}| \geq b) \sim \frac{2mb\varphi(b)}{2\pi} \int_{h_0}^{h_1} \sqrt{\det(\beta_h)} dh$$

where β_h is the covariance matrix of $\partial Z_{t,h}/\partial h$ and $\partial Z_{t,h}/\partial t$. It is tedious but straightforward to compute that

$$\beta_h = \begin{bmatrix} d_{11}(h) & 0 \\ 0 & d_{22}(h) \end{bmatrix}.$$

Substituting β_h into the above expression yields (7).

Appendix C

In this appendix, we prove Theorem 1. Theorem 2 follows from the same arguments and therefore its proof is omitted. The claims stated in the proof will be proved below. At the end of the appendix, we make a comment about local power, which uses similar concepts, although the proof is much easier. For ease of reference, we restate the theorems.

Theorem 1. Let $\tau = \{\tau_1, \dots, \tau_M\}$ and $\mu = \{\mu_1, \dots, \mu_{M+1}\}$ be defined as above; and let $T_{\tau, \mu}$ be defined as in (10). Define $\delta_k = \mu_{k+1} - \mu_k$ for $1 \leq k \leq M$ and $m_k = \tau_k - \tau_{k-1}$ for $1 \leq k \leq M+1$. Suppose that $|\delta_k|$ is bounded away from 0 and ∞ . Suppose further that

$$1 \ll b \ll n_k \ll (m_k \wedge m_{k+1})/b. \quad (22)$$

Let $U_{t, \mu}$ denote the expression on the RHS(10) before taking the maximum. We have,

$$\mathbb{P}_{\tau, \mu}(\max_{t: |t_k - \tau_k| \leq n_k} U_{t, \mu} > b) \sim \mathbb{P}(\sum_{k=1}^M W_k + \frac{1}{2} \chi_{M+1}^2 > b), \quad (23)$$

where $W_1, \dots, W_M, \chi_{M+1}^2$ are independent, χ_{M+1}^2 is a chi-squared random variable with $M+1$ degrees of freedom, and W_k satisfies

$$\mathbb{P}(W_k \geq x) = 2\nu(|\delta_k|)e^{-x} - \nu^2(|\delta_k|)e^{-2x} \quad (24)$$

for $1 \leq k \leq M$.

Theorem 2. Let T_τ be defined as in (12) with the maximum taken over $|t_k - \tau_k| \leq n_k$ for $1 \leq k \leq M$. Define $\hat{\delta}_k = \hat{\mu}_{k+1} - \hat{\mu}_k$ for $1 \leq k \leq M$, $\hat{\mu}_k = (S_{\tau_k} - S_{\tau_{k-1}})/(\tau_k - \tau_{k-1})$ and $m_k = \tau_k - \tau_{k-1}$ for $1 \leq k \leq M+1$. Suppose that $|\hat{\delta}_k|$ is bounded away from 0 and ∞ . Suppose further that

$$1 \ll b \ll n_k \ll (m_k \wedge m_{k+1})/b.$$

We have

$$\mathbb{P}_\tau(T_\tau > b | S_{\tau_1} = s_{\tau_1}, \dots, S_{\tau_M} = s_{\tau_M}) \sim \mathbb{P}(\sum_{k=1}^M W_k > b)$$

where W_1, \dots, W_M are independent have the same distributions as in Theorem 1.

We use C and c to denote positive constants, which may differ in different expressions.

We denote the probability on the right-hand side of (23) by p . Note that

$$p \geq P(W_1 > b) \geq ce^{-b}.$$

We can decompose $U_{t,\mu}$ as

$$U_{t,\mu} = \sum_{k=1}^{M+1} (V_k + Y_k),$$

where

$$V_k = \frac{(S_{t_k} - S_{t_{k-1}})^2}{2(t_k - t_{k-1})} - \frac{(S_{\tau_k} - S_{\tau_{k-1}})^2}{2(\tau_k - \tau_{k-1})}$$

and

$$Y_k = \frac{(S_{\tau_k} - S_{\tau_{k-1}} - (\tau_k - \tau_{k-1})\mu_k)^2}{2(\tau_k - \tau_{k-1})}.$$

Given τ and μ , $\{2Y_k : 1 \leq k \leq M+1\}$ are independent and identically distributed $\chi^2(1)$ random variables. We have

$$\begin{aligned} & \mathbb{P}(\max_{t: |t_k - \tau_k| \leq n_k} U_{t,\mu} > b | \tau, \mu) \\ &= \int_0^{2b} \mathbb{P}(\max_{t: |t_k - \tau_k| \leq n_k} \sum_{k=1}^{M+1} V_k > b - \frac{y}{2} \mid \sum_{k=1}^{M+1} Y_k = \frac{y}{2}, \tau, \theta) f_{\chi_{M+1}^2}(y) dy \\ &+ \int_{2b}^{\infty} f_{\chi_{M+1}^2}(y) dy, \end{aligned} \quad (25)$$

where $f_{\chi_j^2}(\cdot)$ denotes the density function of a $\chi^2(j)$ random variable. Under condition (22), $Y_k \leq b$ implies that

$$\frac{S_{\tau_k} - S_{\tau_{k-1}}}{\tau_k - \tau_{k-1}} = \mu_k + o(1), \quad 1 \leq k \leq M+1. \quad (26)$$

Claim 1. Conditioning on $\{S_{\tau_k} : 1 \leq k \leq M+1\}$ such that (26) is satisfied, we have, with probability $1 - o(p)$,

$$\frac{S_{t_k} - S_{t_{k-1}}}{t_k - t_{k-1}} = \mu_k + o(1), \quad 1 \leq k \leq M+1 \quad (27)$$

for all t such that $|t_k - \tau_k| \leq n_k$.

From Claim 1, in the following we can assume (27). For $|t_1 - \tau_1| \leq n_1$, we have

$$\begin{aligned} \frac{S_{t_1}^2}{2t_1} - \frac{S_{\tau_1}^2}{2\tau_1} &= \frac{1}{2}[(S_{t_1}/t_1)(t_1/\tau_1)^{1/2} + S_{\tau_1}/\tau_1][S_{t_1}(\tau_1/t_1)^{1/2} - S_{\tau_1}] \\ &= (\mu_1 + o(1))[S_{t_1} - S_{\tau_1} - (t_1 - \tau_1)\frac{\mu_1 + o(1)}{2}]. \end{aligned}$$

Similarly, for $2 \leq k \leq M$,

$$\begin{aligned} &\frac{(S_{t_k} - S_{t_{k-1}})^2}{2(t_k - t_{k-1})} - \frac{(S_{\tau_k} - S_{\tau_{k-1}})^2}{2(\tau_k - \tau_{k-1})} \\ &= (\mu_k + o(1))[S_{\tau_{k-1}} - S_{t_{k-1}} + (t_{k-1} - \tau_{k-1})\frac{\mu_k + o(1)}{2}] \\ &\quad + (\mu_k + o(1))[S_{t_k} - S_{\tau_k} - (t_k - \tau_k)\frac{\mu_k + o(1)}{2}], \end{aligned}$$

and

$$\begin{aligned} &\frac{(S_m - S_{t_M})^2}{2(m - t_M)} - \frac{(S_m - S_{\tau_M})^2}{2(m - \tau_M)} \\ &= (\mu_{M+1} + o(1))[S_{\tau_M} - S_{t_M} + (t_M - \tau_M)\frac{\mu_{M+1} + o(1)}{2}]. \end{aligned}$$

Therefore,

$$\max_{t: |t_k - \tau_k| \leq n_k} \sum_{k=1}^{M+1} V_k = \max_{t: |t_k - \tau_k| \leq n_k} \sum_{k=1}^M (-\delta_k)(1 + o(1))[S_{t_k} - S_{\tau_k} - (t_k - \tau_k)\frac{\mu_k + \mu_{k+1} + o(1)}{2}]. \quad (28)$$

Claim 2. Suppose $S_{\tau_k} - S_{\tau_{k-1}}$ satisfies (26). Then, with probability $1 - o(p)$, $\{X_i : \tau_{k-1} < i \leq \tau_{k-1} + n_{k-1}\}$ and $\{X_i : \tau_k - n_k < i \leq \tau_k\}$ are asymptotically independent and identically distributed with distribution $N(\mu_k, 1)$.

Let $\{\xi_j\}_{j \geq 1}$ are independent and identically distributed as $N(0, 1)$. From $n_k \delta_k^2 \gg 1$ and $p \geq ce^{-b}$, we have

$$\sum_{n > n_k} P(|\delta_k|(\sum_{j=1}^n \xi_j - n|\delta_k|) \geq b) \leq C \sum_{n > n_k} e^{-b - \frac{1}{2}n\delta_k^2} = o(p).$$

From (28), Claim 2 and the above bound, we have

$$P(\max_t \sum_{k=1}^{M+1} V_k > b - y/2) \sim P(\sum_{k=1}^M \widetilde{W}_k > b - y/2)$$

where $\{\widetilde{W}_k\}_{1 \leq k \leq M}$ are independent,

$$\widetilde{W}_k = \max\{\widetilde{W}_k^-, \widetilde{W}_k^+\},$$

\widetilde{W}_k^- and \widetilde{W}_k^+ are independent and identically distributed, and

$$\widetilde{W}_k^- = \sup_{i>0} |\delta_k| \left(\sum_{j=1}^i \xi_j - i|\delta_k| \right). \quad (29)$$

We choose y and z such that $1 \ll z \ll \log(y) \ll \log \log(b)$. From (8.49) of Siegmund (1985), we have

$$P(\widetilde{W}_k > z) \sim P(W_k > z).$$

This, together with Claim 3 and Claim 4, proves the theorem.

Claim 3. We have

$$\int_{2b-y}^{\infty} f_{\chi_{M+1}^2}(y) dy = o(p).$$

Claim 4. We have

$$P\left(\sum_{k=1}^M \widetilde{W}_k > y\right) \sim P\left(\sum_{k=1}^M \widetilde{W}_k > y, \min_{1 \leq k \leq M} \widetilde{W}_k > z\right).$$

□

Proof of Claim 1. We only prove for the case $k = 2$. The other cases follow from the same argument. It suffices to show that there exists $a \rightarrow 0$ such that

$$n_1 n_2 P\left(\frac{|S_{t_2} - S_{t_1}|}{t_2 - t_1} \geq a \mid S_{\tau_2} - S_{\tau_1} = 0\right) = o(p) \quad (30)$$

for all t_1, t_2 such that $|t_1 - \tau_1| \leq n_1$ and $|t_2 - \tau_2| \leq n_2$. Note that conditioning on $S_{\tau_2} - S_{\tau_1} = 0$, the mean value of $(S_{t_2} - S_{t_1})/(t_2 - t_1)$ is 0 and by $n_k \ll (m_k \wedge m_{k+1})$, the variance is bounded by $C(n_1 + n_2)/m_2^2$. Therefore,

$$\begin{aligned} & P\left(\frac{|S_{t_2} - S_{t_1}|}{t_2 - t_1} \geq a \mid S_{\tau_2} - S_{\tau_1} = 0\right) \\ & \leq \frac{C\sqrt{(n_1 + n_2)}}{am_2} \exp[-a^2 m_2^2 / (2C(n_1 + n_2)) + \log(n_1) + \log(n_2)]. \end{aligned} \quad (31)$$

For this to be of smaller order than p , we need to choose a such that

$$\frac{a^2 m_2^2}{2C(n_1 + n_2)} - \log(n_1) - \log(n_2) - b \rightarrow \infty.$$

Such an $a \rightarrow 0$ exists because $\frac{m_2^2}{n_1+n_2} \gg b$. \square

Proof of Claim 2. We only prove for the case $k = 2$. The other cases follow from the same argument. Choose D such that

$$D^2 \ll m_2 \text{ and } \frac{D^2}{n_1 + n_2} \gg b.$$

Such a D exists because of (22). Define

$$\tilde{S}'_{n_1} = \sum_{i=\tau_1+1}^{\tau_1+n_1} X_i \text{ and } \tilde{S}''_{n_2} = \sum_{i=\tau_2-n_2+1}^{\tau_2} X_i.$$

By straightforward calculations, we have

$$P(|\tilde{S}'_{n_1} + \tilde{S}''_{n_2} - \mu_2(n_1 + n_2)| \geq D | S_{\tau_2} - S_{\tau_1} = \mu_2) \ll p,$$

and conditioning on the complement event that $|\tilde{S}'_{n_1} + \tilde{S}''_{n_2} - \mu_2(n_1 + n_2)| < D$, $\{X_i : \tau_1 < i \leq \tau_1 + n_1\}$ and $\{X_i : \tau_2 - n_2 < i \leq \tau_2\}$ are asymptotically independent and identically distributed with distribution $N(\mu_2, 1)$. \square

Proof of Claim 3. Note from (24) that the pdf of W_1 for large x is larger than that of $\chi_1^2/2$. Therefore,

$$p \geq cP(\chi_{M+2}^2 \geq 2b) \geq cb^{1/2}P(\chi_{M+1}^2 \geq 2b),$$

which is of higher order than

$$P(\chi_{M+1}^2 \geq 2b - y)$$

by the choice of y . \square

Proof of Claim 4. It suffices to show that for $M \geq 2$,

$$P(\min_{1 \leq k \leq M} \widetilde{W}_k \leq z | \sum_{k=1}^M \widetilde{W}_k > y) = o(1).$$

Note that $P(\sum_{k=1}^M \widetilde{W}_k > y) \asymp y^{M-1}e^{-y}$. Therefore,

$$\begin{aligned} & P(\min_{1 \leq k \leq M} \widetilde{W}_k \leq z | \sum_{k=1}^M \widetilde{W}_k > y) \\ & \leq \frac{MP(\sum_{k=1}^{M-1} \widetilde{W}_k > y - z)}{P(\sum_{k=1}^M \widetilde{W}_k > y)} \asymp \frac{My^{M-2}e^{-y+z}}{y^{M-1}e^{-y}} = o(1) \end{aligned}$$

by the choice of z . □

Local Power When the size of a change in the mean value is δ and the (largest possible) background is (i^*, k^*) for a change-point at j^* , we have defined the marginal power to be $1 - \Phi(b - \delta[h_1 h_2 / (h_1 + h_2)]^{1/2})$, where $h_1 = j^* - i^*$, $h_2 = k^* - i^*$. This is just the marginal probability that the statistic Z evaluated at the true change-point with the largest possible background exceeds the threshold b . Here we consider the probability that a detection occurs due to a local perturbation of the values i^*, j^*, k^* . Let W_0 have the probability distribution $\mathbb{P}\{W_0 \geq x\} = 2\nu(\delta) \exp(-x) - \nu^2(\delta) \exp(-2x)$ for $x \geq 0$. For $i = 1, 2$, let W_i have the distribution given by $\mathbb{P}\{W_i \geq x\} = \nu(\delta_i) \exp(-x)$, where $\delta_1 = \delta/(1 + h_1/h_2)$ and $\delta_2 = \delta/(1 + h_2/h_1)$. The local power is then the sum of the marginal power and the perturbation

$$2 \int_0^{b^2/2} \mathbb{P}\{\sum W_i > b^2/2 - x\} f(2x; 1, \lambda) dx,$$

where f is the probability density function of a χ^2 distribution with one degree of freedom and noncentrality parameter $\lambda = \delta^2 h_1 h_2 / (h_1 + h_2)$.

Acknowledgement. The authors thank Nancy Zhang for several helpful discussions and suggestions.

XF was partially supported by the National University of Singapore grant R-155-000-158-112. DS was partially supported by the National Science Foundation.

References

- R. J. Adler and J. E. Taylor *Random Fields and Geometry* Springer-Verlag, New York-Heidelberg-Berlin.
- G. A. Churchill (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79–94.
- P. Diaconis and D. A. Freedman (1988). Conditional limit theorems for exponential families and finite versions of de Finetti’s theorem. *J. Theoret. Probab.* **1**, 381–410.
- C. Du, C. L. M. Kao and S. C. Kou (2016). Stepwise signal extraction via marginal likelihood. *J. Amer. Statist. Assoc.* **111**, 314–330.
- E. Elhaik, D. Graur and K. Josić (2010) Comparative testing of DNA segmentation algorithms using benchmark simulations, *Mol. Bio. Evol.* **27** 1015-1024.
- X. Fang and D. O. Siegmund (2015). Multiscale methods to detect local signals: Strengths and weaknesses *Unpublished manuscript*.

- K. Frick, A. Munk and H. Sieling (2014). Multiscale change point inference. With 32 discussions by 47 authors and a rejoinder by the authors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76**, 495–580.
- P. Fryzlewicz (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42**, 2243–2281.
- W. R. Lai, M. D. Johnson, R. Kucherlapati and P. J. Park (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763–3770.
- Y.S. Niu and H. Zhang (2012). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Statist.* **6**, 1306–1326.
- A. B. Olshen, E. S. Venkatraman, R. Lucito and M. Wigler (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572.
- F. Picard, S. Robin, M. Lavielle, C. Vaisse and J. J. Daudin (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6**, 27.
- J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays, *Nat. Genet.* **23**, 41–46.
- J. R. Pollack, T. Sørli, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Børresen-Dale, and P. O. Brown (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors., *Proc. Natl. Acad. Sci. USA* **99**, 12963–12968.
- D. O. Siegmund (1985). *Sequential analysis: Tests and confidence intervals*. Springer, New York.
- D. O. Siegmund (1988). Confidence sets in change-point problems. *Internat. Statist. Rev.* **56**, 31–48.
- D. O. Siegmund and B. Yakir (2007). *The Statistics of Gene Mapping*. Springer, New York.
- A. M. Snijders, J. Fridlyand, D. A. Mans, R. Segreaves, A. N. Jain, D. Pinkel, and D. G. Albertson (2003). Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene* **22**, 4370–4379.
- L. Vostrikova (1981). Detecting ‘disorder’ in multidimensional random processes. *Soviet Math. Dokl.* **24**, 55–59.
- B. Yakir (2013). *Extremes in random fields : a theory and its applications*. Wiley, United Kingdom.
- K. J. Worsley (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* **73**, 91–104.
- N. R. Zhang and D. O. Siegmund (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data *Biometrics* **63**, 22–32.

- N. R. Zhang, D. O. Siegmund, H. Ji and J. Z. Li (2010). Detecting simultaneous changepoints in multiple sequences. With supplementary data available online. *Biometrika* **97**, 631–645.
- Y. Zhang and J. Liu (2011). Fast and accurate approximation to significance tests in genome-wide association studies. *J. Amer. Statist. Assoc.* **106**, 846–857.
- X. Zhao, C. Li, J. G. Paez, K. Chin, P. A. Jänne, T.-H. Chen, L. Girard, J. Minna, D. Christiani, C. Leo, J. W. Gray, W. R. Sellers, and M. Meyerson (2004). An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays *Cancer Res.* **64**, 3060–3071.